Revised Manuscript (unmarked)
Click here to view linked References

A spline-based tool to assess and visualize the calibration of multiclass risk predictions

Van Hoorde K[a,b], Van Huffel S[a,b], Timmerman D[c,d], Bourne T[c,d,e], Van Calster B[d,f]

Affiliations:

[a] KU Leuven Department of Electrical Engineering (ESAT) STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Leuven, Belgium

[b] KU Leuven iMinds Medical Information Technologies, Leuven, Belgium

[c] Department of Obstetrics & Gynecology, University Hospitals Leuven, Leuven, Belgium

[d] KU Leuven Department of Development & Regeneration, Leuven, Belgium

[e] Queen Charlotte's & Chelsea Hospital, Imperial College, Du Cane Road, London, W12 0HS, UK

[f] Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

Corresponding author:

Ben Van Calster

KU Leuven Department of Development & Regeneration,

Herestraat 49 Box 7003

B3000 Leuven, Belgium

E-mail address: ben.vancalster@med.kuleuven.be

Running title:

Calibration of multiclass risk prediction models

Word count:

4296 words

Keywords (max 6):

risk models; probability estimation; machine learning; logistic regression; calibration; multiclass

Abstract

245 words

When validating risk models (or probabilistic classifiers), calibration is often overlooked. Calibration refers to the reliability of the predicted risks, i.e. whether the predicted risks correspond to observed probabilities. In medical applications this is important because treatment decisions often rely on the estimated risk of disease. The aim of this paper is to present generic tools to assess the calibration of multiclass risk models.

We describe a calibration framework based on a vector spline multinomial logistic regression model. This framework can be used to generate calibration plots and calculate the estimated calibration index (ECI) to quantify lack of calibration. We illustrate these tools in relation to risk models used to characterize ovarian tumors. The outcome of the study is the surgical stage of the tumor when relevant and the final histological outcome, which is divided into five classes: benign, borderline malignant, stage I, stage II-IV, and secondary metastatic cancer. The 5909 patients included in the study are randomly split into equally large training and test sets. We developed and tested models using the following algorithms: logistic regression, support vector machines, k nearest neighbors, random forest, naive Bayes and nearest shrunken centroids.

Multiclass calibration plots are interesting as an approach to visualizing the reliability of predicted risks. The ECI is a convenient tool for comparing models, but is less informative and interpretable than calibration plots. In our case study, logistic regression and random forest showed the highest degree of calibration, and the naive Bayes the lowest.

Abbreviations

AUC	Area under the ROC curve
CI	Confidence Interval
ECI	Estimated Calibration Index
IOTA	International Ovarian Tumor Analysis
IQR	InterQuartile Range
kNN	k Nearest Neighbor
LR-BT	Logistic Regression Binary Tree – sequential dichotomous model
LR-PC	Logistic Regression Pairwise Coupling
MLR	Multinomial Logistic Regression
NB	Naive Bayes
NSC	Nearest Shrunken Centroids
PDI	Polytomous Discrimination Index
RF	Random Forest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machines
SVM-BT	Support Vector Machines Binary Tree – sequential dichotomous model
SVM-PC	Support Vector Machines Pairwise Coupling

1 Introduction

For medical applications, prediction models that provide probabilistic (risk) estimates of an event of interest are useful for clinical decision support, personalized healthcare, and shared decision making. Prior to the implementation of such tools in clinical practice, validation with respect to discrimination and calibration is required [1-6]. A model needs to be able to distinguish between different possible outcome categories (discrimination). How well a model performs this function can be evaluated using the area under the receiver operating characteristic curve (AUC) or multiclass extensions of this approach. Calibration assessment is often overlooked, but is of importance for several applications where risk models may be used. Such applications include decisions whether or not to treat a patient [7] start preventive action, or to inform the choice of treatment [8] Calibration is also relevant when informing patients about risk [9], when comparing hospitals with respect to quality of care (e.g. benchmarking based on mortality risk) [10], and when identifying high risk patients for inclusion in clinical trials [11]. The optimal use of risk models in these situations relies on reliable risk estimation. For example, a classic result from decision analysis states that the adopted risk threshold to decide whether or not to take further action implies there are relative misclassification costs [12]: the odds of the risk threshold equals the ratio of the harm of a false positive test result to the benefit of a true positive result. For example if a risk threshold of 10% is adopted, the assumption is that 1 true positive is worth 9 false positives. If a poorly calibrated risk model is then used to assess whether patients exceed the planned threshold, inappropriate decisions may be taken.

For binary outcomes, the relationship between predicted and observed probabilities can be visualized by means of a calibration plot [1, 13, 14]. Observed probabilities are sometimes obtained by computing event rates within groups of patients with similar predicted probabilities (e.g. decile split). However sometimes, flexible smoothing methods such as local regression (loess) or splines are used to link predicted probabilities to estimated observed probabilities [1].

Recently, our group extended binary calibration plots to multiclass models based on multinomial logistic regression (MLR) [15]. For calibration statistics we introduced the concept of multiclass calibration-in-the-large [15]. We proposed two frameworks, one parametric and one non-parametric. Logistic regression is a common algorithm to build binary and multiclass clinical prediction models, and naturally works with risk estimates. However, machine learning algorithms are also used for clinical risk prediction [16-21], and are very frequently used in high dimensional and/or “large p, small n” prediction studies (i.e. a large number of predictors and a small number of patients) [22-24]. Moreover, although using machine-learning approaches for classification problems is often less suited to probability estimation, methods do exist to facilitate this [25-30]. The calibration performance of risk models is also an issue that is often neglected, and it is not surprising that with a few exceptions this is frequently the case for models based on machine learning algorithms [13, 26, 27, 30-32].

The aim of this paper is to introduce a non-parametric framework to evaluate the calibration of multiclass risk models irrespective of the modeling technique used. Based on this framework we also derive a calibration measure to quantify and compare calibration performance between models. We illustrate these methods with a case study looking at the classification of ovarian tumors. We develop and validate risk models to diagnose tumor pathology based on logistic regression, support vector machines, k-nearest neighbors, random forest, naive Bayes and nearest shrunken centroids.

2 Non-parametric recalibration framework

Our group developed calibration tools for risk models based on multinomial logistic regression (MLR) [15]. Assume an MLR or 'baseline-category logit' model [33] with *m* predictors x_1 to x_m for an outcome with *J* (*j=1, …, J*) categories. If category 1 is chosen as the reference category, the model is written as

$$\begin{cases} \log\left[\dfrac{P(Y=2)}{P(Y=1)}\right] = \alpha_2 + \sum_{l=1}^{m} \beta_{2l}\, x_l = \mathrm{lp}_{21} \\ \log\left[\dfrac{P(Y=3)}{P(Y=1)}\right] = \alpha_3 + \sum_{l=1}^{m} \beta_{3l}\, x_l = \mathrm{lp}_{31} \\ \ldots \\ \log\left[\dfrac{P(Y=J)}{P(Y=1)}\right] = \alpha_J + \sum_{l=1}^{m} \beta_{Jl}\, x_l = \mathrm{lp}_{J1} \end{cases} \tag{1}$$

and the multiclass risks are obtained as

$$\begin{cases} P(Y=1) = p_1 = \dfrac{1}{1+\exp(\mathrm{lp}_{21})+\exp(\mathrm{lp}_{31})+\ldots+\exp(\mathrm{lp}_{J1})} = \dfrac{1}{1+\sum_{j=2}^{J}\exp(\mathrm{lp}_{j1})} \\ P(Y=2) = p_2 = \dfrac{\exp(\mathrm{lp}_{21})}{1+\exp(\mathrm{lp}_{21})+\exp(\mathrm{lp}_{31})+\ldots+\exp(\mathrm{lp}_{J1})} = \dfrac{\exp(\mathrm{lp}_{21})}{1+\sum_{j=2}^{J}\exp(\mathrm{lp}_{j1})} \\ \ldots \\ P(Y=J) = p_J = \dfrac{\exp(\mathrm{lp}_{J1})}{1+\exp(\mathrm{lp}_{21})+\exp(\mathrm{lp}_{31})+\ldots+\exp(\mathrm{lp}_{J1})} = \dfrac{\exp(\mathrm{lp}_{J1})}{1+\sum_{j=2}^{J}\exp(\mathrm{lp}_{j1})} \end{cases} \tag{2}$$

Let $\{\widehat{\mathrm{lp}}_{21},\ldots,\widehat{\mathrm{lp}}_{J1}\}$ denote the estimated linear predictors and $\{\hat{p}_1,\ldots,\hat{p}_J\}$ the estimated multiclass risks. The non-parametric recalibration framework for such models relates the multiclass outcome *Y* on the estimated *J-1* linear predictors $\{\widehat{\mathrm{lp}}_{21},\ldots,\widehat{\mathrm{lp}}_{J1}\}$ from the MLR risk model through a vector spline [34] MLR analysis [15]:

$$\begin{cases}\log\left[P(Y=2)/P(Y=1)\right]=a_2+\sum_{j=2}^{J}\left(b_{2j}\cdot s_2\left(\hat{\mathrm{lp}}_{j1}\right)\right)\\ \log\left[P(Y=3)/P(Y=1)\right]=a_3+\sum_{j=2}^{J}\left(b_{3j}\cdot s_3\left(\hat{\mathrm{lp}}_{j1}\right)\right)\\ \ldots\\ \log\left[P(Y=J)/P(Y=1)\right]=a_J+\sum_{j=2}^{J}\left(b_{Jj}\cdot s_J\left(\hat{\mathrm{lp}}_{j1}\right)\right)\end{cases} \tag{3}$$

with $s(.)=\left[s_2(.)\ s_3(.)\ \ldots\ s_J(.)\right]$ a vector spline smoother applied to each linear predictor [15, 34]. This vector spline smoother $s(.)$ is a natural extension of the cubic spline smoother to vector responses and consists of J-*1* natural cubic B-splines $s_j(.)$ [34, 35]. Similarly as the multiclass risks in (2) are obtained from (1), this framework can be used to estimate the observed probabilities $\{\hat{o}_1,\ldots,\hat{o}_J\}$ (Appendix A) [15].

If we now consider a generic multiclass risk model (i.e. irrespective of how it was developed) yielding risk estimates $\{\hat{p}_1,\ldots,\hat{p}_J\}$. In order to use the nonparametric recalibration model we first calculate $\log\left(\hat{p}_j/\hat{p}_1\right)=\hat{\mathrm{z}}_{j1}$, with *j=2, …, J* if category 1 is used as reference category. These quantities are used as predictors in the vector spline MLR model (3) to estimate the observed probabilities:

$$\begin{cases}\log\left[P(Y=2)/P(Y=1)\right]=a_2+\sum_{j=2}^{J}b_{2j}\cdot s_2\left(\hat{z}_{j1}\right)\\ \log\left[P(Y=3)/P(Y=1)\right]=a_3+\sum_{j=2}^{J}b_{3j}\cdot s_3\left(\hat{z}_{j1}\right)\\ \ldots\\ \log\left[P(Y=J)/P(Y=1)\right]=a_J+\sum_{j=2}^{J}b_{Jj}\cdot s_J\left(\hat{z}_{j1}\right)\end{cases} \tag{4}$$

The observed probabilities $\hat{o}_j$ are obtained by using the vector spline MLR model (4) and then applying the equations like (2) (Appendix A). Calibration plots are obtained by plotting the predicted probabilities $\hat{p}_j$ versus the observed probabilities $\hat{o}_j$ for each outcome category *j* (*j=1*,…,*J*) [15]. Note that there is no one-to-one relationship between $\hat{p}_j$ and $\hat{o}_j$. The reason is that, for a specific value of $\hat{p}_j$, predicted probabilities for the other J-1 categories can vary and this will lead to different values for $\hat{o}_j$. Therefore, spline smoothers are plotted to assess the trend of the scatter plot for each category [15].

To quantify the lack of calibration with a single measure, the correspondence between $\hat{p}_j$ and $\hat{o}_j$ can be assessed through their squared difference averaged over J categories and N observations: $\left[\sum_{n=1}^{N}\sum_{j=1}^{J}\left(\hat{p}_{nj}-\hat{o}_{nj}\right)^2\right]\Big/\left(N\cdot J\right)$. If we multiply the average squared difference by 100J/2 we obtain a measure that we refer to as the Estimated Calibration Index (ECI) because it is based on estimates of the observed probabilities $\left\{\hat{o}_1,\ldots,\hat{o}_J\right\}$. The multiplication factor 100J/2 ensures that the ECI has a theoretical range between 0 and 100 (Appendix B). The ECI has similarities to the Brier score. However, the Brier score represents the average squared difference between the actual outcomes y_n and the predicted probabilities such that it is an overall performance measure that captures discrimination and calibration: Brier score = $\left[\sum_{n=1}^{N}\sum_{j=1}^{J}\left(y_{nj}-\hat{p}_{nj}\right)^2\right]\Big/\left(N\cdot J\right)$ [36]. In contrast, the ECI is the average squared difference of the predicted probabilities $\hat{p}_n$ with the estimated observed probabilities $\hat{o}_n$ instead of the actual outcomes.

3 Case-study

The accurate diagnosis of ovarian tumors prior to surgery is crucial when choosing appropriate patient management and referral. As different types of malignancies require different management, we aimed to develop a model to predict the risk that an ovarian tumor is benign, borderline malignant, a stage I cancer, a stage II-IV cancer, or secondary metastatic cancer. We used data from the International Ovarian Tumor Analysis (IOTA) consortium [37-39]. These data are derived from 5909 patients collected between 1999 and 2012 at 24 centers in 10 countries [40], and were randomly divided into a development and validation set stratified for the final outcome. The outcome is based on the histo-pathological diagnosis of the mass after surgical removal by laparotomy or laparoscopy as well as the stage of the tumor (when relevant) as assessed by the surgeon. Pathologists were blind to the data collected for the study and the results of any risk models we developed. The following predictors or features were considered for model development without further selection: age (years), serum CA125 (U/ml), oncology referral center (yes/no), maximum diameter of the lesion (mm), proportion of solid tissue (between 0 and 1), presence of more than 10 cyst locules (yes/no), number of papillary projections (0, 1, 2, 3 and more than 3), presence of acoustic shadows (yes/no) and presence of ascites (yes/no) (Table 1). The proportion of solid tissue is defined as the ratio of the maximum diameter of the largest solid component and the maximum diameter of the lesion. These nine variables are the predictors from the multinomial risk prediction model ADNEX [40]. Some values for serum CA125 were missing and were imputed in order not to lose these records. We used predictive mean matching regression [41] using variables that were either related to the level of CA125 itself, or the unavailability of CA125 via an indicator as to whether the CA125 level was missing or not [42].

Table 1. Descriptive statistics of the ovarian tumor case study

	Benign	Borderline	Stage I	Stage II-IV	Metastatic
Outcome, N	3980	339	356	988	246
Variable, N (%) or median (IQR)					
Age (years)	42 (32-54)	49 (36-62)	54 (44-64)	59 (50-67)	57 (47-68)
Serum CA125 (U/mL)*	19 (11-39)	31 (16-100)	52 (21-190)	447 (147-1215)	81 (30-271)
Family history of ovarian cancer,	79 (2.0)	10 (3.0)	13 (3.7)	57 (5.8)	5 (2.0)
Maximal diameter of lesion (mm)	63 (45-87)	86 (51-150)	106 (71-153)	85 (56-123)	86 (56-124)
Solid tissue					
Presence of solid tissue	1322 (33.2)	267 (78.8)	328 (92.1)	968 (98.0)	234 (95.1)
Proportion solid tissue if present (%)	42 (20-100)	37 (24-59)	61 (38-100)	100 (56-100)	100 (64-100)
Number of papillary projections					
None	3424 (86.0)	135 (39.8)	227 (63.8)	772 (78.1)	213 (86.6)
1	333 (8.4)	69 (20.4)	25 (7.0)	56 (5.7)	12 (4.9)
2	80 (2.0)	21 (6.2)	17 (4.8)	30 (3.0)	0 (0)
3	66 (1.7)	24 (7.1)	17 (4.8)	28 (2.8)	2 (0.8)
>3	77 (1.9)	90 (26.5)	70 (19.7)	102 (10.3)	19 (7.7)
More than 10 cyst locules	199 (5.0)	74 (21.8)	69 (19.4)	93 (9.4)	36 (14.6)
Acoustic shadows	676 (17.0)	8 (2.4)	18 (5.1)	30 (3.0)	10 (4.1)
Ascites	64 (1.6)	28 (8.3)	65 (18.3)	473 (47.9)	90 (36.6)
Missing values for CA125, N (%)	1447 (36.4)	62 (18.3)	71 (19.9)	163 (16.5)	62 (25.2)

* Results for Serum CA125 are based on single imputation of missing values.

Abbreviations: IQR; Interquartile Range.

3.1 Study set-up

We develop and validate various binary and multiclass risk models to diagnose ovarian tumors. The models are based on logistic regression and machine learning algorithms (see Appendix B). In the first part of the study, we develop binary models to distinguish between benign and malignant tumors. We then discuss multiclass problems by developing models to distinguish between benign, borderline malignant, stage I invasive, stage II-IV invasive and secondary metastatic ovarian tumors. The available data were split into a development set and a validation set. The split was random, using a 1:1 ratio with stratification according to the multiclass outcome. R version 3.0.3 (www.r-project.org) was used for the statistical analysis.

3.2 Performance evaluation

The ability to distinguish between different outcome categories (discrimination) as well as the reliability of predicted risks (calibration) is assessed.

3.2.1 Discrimination

Discrimination of a binary risk prediction model was evaluated using the area under the ROC curve (AUC) [43]. For the multiclass risk models, several measures were used that are related to the AUC. First, we calculated AUCs for pairs of categories based on conditional risks (e.g. $risk_A/(risk_A+risk_B)$, where $risk_A$ and $risk_B$ are the estimated risks for categories A and B) [44]. Two multiclass AUCs variants were also evaluated: the M-index and Polytomous Discrimination Index (PDI) [45, 46]. The M-index is an average of pairwise AUCs, where the pairwise AUC for categories A and B is the average of the AUC based on $risk_A$ and the AUC of $risk_B$ (these two AUCs are not the same because $risk_A+risk_B \neq 1$) [45]. The PDI is a direct multiclass version of the AUC and estimates the average proportion of correctly identified cases within a group of cases each belonging to a different outcome category [46]. For J outcome categories the value of M-index and PDI for a model that gives random predictions is 0.5 and $1/J$ (e.g. 0.2 if J=5) [46].

3.2.2 Calibration

As an overall group-level measure we calculated the calibration-in-the-large or mean calibration for every outcome category. This measure equates to the difference between the observed event rate of an outcome category and the average predicted probability of this category (i.e. the predicted event rate). For the assessment of the calibration of individual risk predictions we used the non-parametric recalibration framework explained in Section 2 to produce calibration plots and to calculate ECI.

3.3 Overview of the implemented approaches

In the first part of the study we focus on the binary diagnosis of ovarian tumors as benign or malignant. The binary risk models were logistic regression, support vector machines, k nearest neighbors, random forest, naive bayes and nearest shrunken centroids [47-51] (Table 2). More information concerning the different models and the tuning of the hyperparameters is given in Appendix C.

In the second part we focus on the multiclass diagnosis of ovarian tumors as benign, borderline malignant, stage I cancer, stage II-IV cancer, or secondary metastatic cancer. Multiclass risk models can either be a combination of binary risk models or an 'all-in-one' model. Among the algorithms we implemented, only the SVMs could not give all-in-one multiclass risks. When binary models were combined, two types of combinations were used: a tree of nested dichotomies (or binary tree) [52], and pairwise coupling [53]. For the former a tree of binary prediction models is constructed. In our case study, from a clinical viewpoint the first two dichotomies of the binary tree are straightforward: first a model to distinguish between benign and malignant tumors, then for the malignant tumors a model to distinguish between borderline and invasive tumors. The invasive tumor category has three subgroups: stage I, stage II-IV and secondary metastatic cancers. Since it is less obvious what the most clinically relevant tree is for these three categories, we constructed three possible trees and computed the average (Figure 1). For pairwise coupling, the pairwise probabilities $\hat{p}_{ij}$ of every pairwise model *i* versus *j* are combined in order to obtain multiclass probabilities $\hat{p}_i$ ($i=1,\ldots,J$) by solving the following system: $p_i = \sum_{j=1, j\neq i}^{J} \left[\left(p_i + p_j \right) / J - 1 \right] \hat{p}_{ij}$, for all *i* with $\sum_{i=1}^{J} p_i = 1$ and $p_i \geq 0$ [54].

Figure 1. The three different considered sequential dichotomous models or binary trees.

For multiclass risk prediction, we applied nine algorithms (Table 2). For logistic regression, all-in-one multinomial logistic regression (MLR), sequential dichotomous logistic regression (LR-BT) and pairwise coupling logistic regression (LR-PC) were used. For support vector machines, a binary tree (SVM-BT) and a pairwise coupling (SVM-PC) approach were applied.

Table 2. Overview of the binary and multiclass approaches considered.

Binary	Multiclass		
	all-at-once	**binary tree**	**pairwise coupling**
LR	MLR	LR-BT	LR-PC
SVM	-	SVM-BT	SVM-PC
KNN	KNN	-	-
RF	RF	-	-
NB	NB	-	-
NSC	NSC	-	-

3.4 Model comparison based on ECI

Confidence intervals for ECI and the ECI difference between two models were obtained with bootstrapping using the bias-corrected method on 1000 bootstrap samples of the validation data. We always reported 95% confidence intervals, however we used the following procedure to test for differences between models with respect to calibration. We ranked the models based on their validation data ECI, and compared the best model with every other model. A correction for multiple testing was used using a method similar to the Holm step-down method. We compared the best model with the worst model using $\alpha=0.05/(m-1)$, where m is the number of models. If significant, the best model is compared with the second to worst model using $\alpha=0.05/(m-2)$, and so on. Once a test is not statistically significant, all remaining tests are considered not significant as well.

3.5 Software

For the logistic regression and machine learning models we used the following R-packages: `rms`, `vgam`, `klaR`, `pamr`, `randomForest`, `kernlab` and `caret`. The function `sigest()` of the R-package kernlab is used for automatic sigma estimation, while the functions `train()` and `trainControl()` of the R-package `caret` were used for the determination of optimal tuning parameters.

4 Results

4.1 Binary outcome

The optimal hyperparameter values based on maximizing the accuracy and minimizing the logloss were different for the SVM regularization parameter *C* (10 based on accuracy vs 1 based on logloss) and for the number of neighbors for kNN (20 vs 50). Minimizing the logloss resulted in a simpler model compared with maximizing accuracy. This implies less over-fitting and smoother results, which means that better calibration performance can be expected.

The validation AUCs demonstrated good discrimination for all approaches (Table 3). kNN had the worst discriminative ability with a validation AUC of 0.890, RF the best with a validation AUC of 0.948.

Table 3. Discrimination (AUC) and calibration (mean calibration (in %) of the risk of malignancy and estimated calibration index (ECI)) performance of the binary prediction models on the validation data.

Model (selection criterion)	**Discrimination**	**Calibration**	
	AUC	**Mean calibration**	**ECI**
LR	0.938	0.25	0.01
SVM (accuracy)	0.940	0.94	0.09
SVM (logloss)	0.945	0.40	0.04
kNN (accuracy)	0.860	1.80	0.61
kNN (logloss)	0.862	1.92	0.05
RF (accuracy)	0.948	1.64	0.05
RF (logloss)	0.948	1.64	0.05
NB (accuracy)	0.910	3.93	1.67
NB (logloss)	0.910	3.93	1.67
NSC (accuracy)	0.890	2.47	0.45
NSC (logloss)	0.890	2.47	0.45

The mean calibration of the risk of malignancy was close to zero for logistic regression. Overall it ranged between 0.25% (logistic regression) and 3.93% (naive Bayes) on the validation data suggesting that the risk of malignancy was on average slightly underestimated (Table 3). Most models showed an ECI close to zero, exceptions based on accuracy were kNN, NB, and NSC (Table 3). Logistic regression had the best ECI. Ranking models based on mean calibration or ECI yields different results, which is explained by the fact that these measures do not capture the same information. In general terms the mean calibration measures calibration on the group level whereas ECI focuses on the individual level. For example, kNN based on logloss has relatively poor mean calibration but a relatively good ECI.

The calibration plots for logistic regression (lowest ECI) and NB (highest ECI) are shown in Figure 2 a-b. We have plotted the curves for the risk of a benign tumor and for the risk of malignancy, even though plotting only one would be sufficient because both curves are complementary. The reason we have shown both curves is to make the plots consistent with the calibration plots for multiclass models, where it is convenient to show curves for every category. It is interesting to observe that the ECI value of 1.67 for NB corresponded to calibration curves that are very deviant from the (ideal) diagonal line. Although 1.67 appears close to 0, the NB model is not well calibrated.

For SVM and kNN, different hyperparameters were selected when relying on logloss vs accuracy as a selection criterion. As expected, models based on logloss showed better calibration. This is illustrated in Figure 2c-d for kNN. A similar yet less explicit difference was observed for SVM.

Figure 2. Calibration plot of the validation data for (a) dichotomous logistic regression, (b) binary naive Bayes, (c) k nearest neighbors with maximizing accuracy and (d) minimizing logloss as selection criterion.

4.2 Multiclass outcome

For all machine learning approaches except RF, hyperparameter values obtained by maximizing accuracy were different from the values obtained by minimizing logloss resulting in different risk probabilities.

The best multiclass AUC results were obtained for the logistic regression models (0.790-0.798) and RF (0.792), the worst for kNN (0.684-0.686) (Table 4). Regarding pairwise AUCs (Table S1), it was clear that it was most difficult to distinguish between borderline and stage I cancer, and between the different types of invasive cancers (stage I, stage II-IV, and secondary metastatic cancer).

Table 4. Discrimination, i.e. M-index and polytomous discrimination index PDI, for the multiclass prediction models using the validation data.

Model (selection criterion)	**Discrimination**		**Calibration**
	M-index	**PDI**	**ECI (95%CI)**
MLR	0.790	0.529	0.27 [0.13;0.36]
LR-BT	0.798	0.542	0.30 [0.18;0.38]
LR-PC	0.795	0.537	0.26 [0.18;0.32]
SVM-BT (acc)	0.755	0.482	0.51 [0.31;0.61]
SVM-BT (ll)	0.770	0.505	0.38 [0.26;0.48]
SVM-PC (acc)	0.764	0.467	0.76 [0.59;0.87]
SVM-PC (ll)	0.764	0.469	0.81 [0.64;0.93]
kNN (acc)	0.684	0.403	0.76 [0.56;0.94]
kNN (ll)	0.686	0.399	0.28 [0.18;0.33]
RF (acc)	0.792	0.539	0.40 [0.23;0.48]
RF (ll)	0.792	0.539	0.40 [0.23;0.48]
NB (acc)	0.742	0.451	8.25 [7.66;8.76]
NB (ll)	0.768	0.495	4.82 [4.26;5.39]
NSC (acc)	0.754	0.470	1.42 [1.11;1.63]
NSC (ll)	0.753	0.471	1.19 [0.91;1.39]

Overall the logistic regression approaches had the best mean calibration, while NB was clearly worse, specifically when accuracy was used to tune the hyperparameter (Table S2). This is remarkable given that the available data were randomly split into development and validation sets. For the other models the deviation was mild (up to 4%). The ECI was lowest for the logistic regression approaches (0.26-0.30) and kNN based on logloss (0.28), and highest for NB (4.82 when based on logloss, 8.25 when based on accuracy) (Table 4). When comparing the ECI of the best model LR-PC with those of every other model, with the application of a Holm-based step-down method to correct for multiple testing, we conclude that there is no statistically significant difference between the ECI of LR-PC and the ECI's of MLR, LR-BT, SVM-BT based on logloss, kNN based on logloss, and RF (Table 5). The other models' ECI was significantly lower.

The multiclass calibration plots of the best calibrated models according to ECI (LR-PC as the best logistic regression model, SVM-BT based on logloss as the best SVM model, RF, and kNN based on logloss) are shown in Figure 3. As highlighted in section 2,, contrary to the calibration plots for binary outcomes, there is no one-to-one relationship between predicted and observed probabilities. For that reason we added spline smoothers to summarize the trend. In Figure 4 simplified calibration plots with only the smoothed curves are shown. All models are fairly well calibrated. The most clear deviation is that higher estimated risks for the small categories (borderline, stage I cancer, secondary metastatic cancer) are typically too extreme (overfitted) as the curves for these outcome categories deviate more from the diagonal as the predicted probability increases.

Table 5. Model comparison based on the difference in ECI (with 95% confidence intervals) using the validation data. In the first part, the ECI of each model is compared with the ECI of the best model (LR-PC). Models for which there is a statistically significant difference with LR-PC after correction for multiple testing are identified with an asterisk. In the second part, the ECI values of each machine learning model with hyperparameter tuning based on accuracy and logloss are compared.

Model comparison	**ECI (95% CI)**
Each model vs the best (LR-PC)	
MLR vs LR-PC	0.01 [-0.05;0.07]
kNN (ll) vs LR-PC	0.02 [-0.10;0.17]
LR-BT vs LR-PC	0.03 [-0.02,0.10]
SVM-BT (ll) vs LR-PC	0.12 [0.01;0.24]
RF (acc) vs LR-PC	0.14 [0.02;0.29]
RF (ll) vs LR-PC	0.14 [0.02;0.29]
SVM-BT (acc) vs LR-PC	0.25 [0.09;0.40]*
kNN (acc) vs LR-PC	0.50 [0.28;0.75]*
SVM-PC (acc) vs LR-PC	0.49 [0.32;0.67]*
SVM-PC (ll) vs LR-PC	0.55 [0.37;0.72]*
NSC (ll) vs LR-PC	0.93 [0.72;1.15]*
NSC (acc) vs LR-PC	1.15 [0.94;1.40]*
NB (ll) vs LR-PC	4.55 [4.06;5.17]*
NB (acc) vs LR-PC	7.99 [7.45;8.59]*
Accuracy versus logloss	
SVM-PC: acc vs ll	-0.05 [-0.14;0.05]
SVM-BT: acc vs ll	0.13 [0.01;0.25]
NSC: acc vs ll	0.22 [0.17;0.29]
kNN: acc vs ll	0.47 [0.33;0.63]
NB: acc vs ll	3.44 [2.59;4.20]

Figure 3. Multiclass calibration plot with smoother using the validation data for (a) pairwise coupling logistic regression model, (b) binary tree support vector machines

with minimizing logloss as selection criterion, (c) random forest and (d) k nearest neighbors with minimizing logloss as selection criterion.

Figure 4. Smoothed multiclass calibration plot using the validation data for (a) pairwise coupling logistic regression model, (b) binary tree support vector machines with minimizing logloss as selection criterion, (c) random forest and (d) k nearest neighbors with minimizing logloss as selection criterion.

The ECI was always lower when hyperparameter tuning was based on logloss compared with accuracy, except for SVM-PC where results were nearly identical for both methods (Table 4). For four algorithms (SVM-BT, kNN, NSC, NB), the 95% confidence interval on the ECI difference did not include 0 (Table 5). Even if we were to correct for multiple testing, significant differences would remain except for SVM-BT. Hence minimization of logloss resulted in better calibration, which is illustrated for kNN and NB in Figure 5. This figure also shows the clear miscalibration of risks based on NB, even when logloss was used.

Figure 5. Smoothed multiclass calibration plot using the validation data for k nearest neighbors (a-b) and naive Bayes (c-d) with maximizing accuracy and minimizing logloss as selection criterion.

5 Discussion

When validating a risk model for medical applications, discrimination as well as calibration should be assessed [1-6]. The model needs to be able to distinguish the different outcome categories but also to reliably predict the risk of a certain outcome. In this paper we have described a generic recalibration framework to visualize and quantify the lack of calibration for multiclass risk models irrespective of the modeling approach used. To illustrate this we have used the diagnosis of ovarian tumors as a case study. Different models were developed and validated with respect to discrimination and calibration: logistic regression, support vector machines, k nearest neighbors, random forest, naive Bayes and nearest shrunken centroids. Overall, logistic regression and random forest models showed the best calibration, whereas the calibration of naive Bayes was disappointing. An explanation for the latter might be that the independence assumption is unrealistic for the data leading to inaccurate estimates [55].

5.1 Visualization versus quantification of calibration

The recalibration framework was used to visualize calibration by means of a calibration plot as well as to quantify the lack of calibration through the estimated calibration index (ECI). The ECI focuses on calibration of individual patient risks, and therefore has an advantage over mean calibration, which is a group level measure. Mean calibration assesses whether estimated event rates are accurate. Lack of mean calibration is also captured by ECI, but ECI captures other aspects as well such as over-fitting (risk estimates that are too extreme).

Both visualization and quantification have their advantages and disadvantages. Calibration plots are more informative and interpretable than ECI, because ECI summarizes the calibration plot for every outcome into a single number. Therefore, for the evaluation of a single model one should focus on calibration plots. The ECI should not be used as a stand-alone metric in this situation because the result is hard to interpret without corresponding plots. The well-known Hosmer-Lemeshow goodness-of-fit test and its extensions are often used to evaluate miscalibration of a single model. Hosmer-Lemeshow statistics were not discussed in this paper due to reported drawbacks with their use such as instability, low power, sensitivity to sample size and arbitrariness of the construction of subgroups [56, 57]. If the aim is to compare multiple models, however, ECI can be used. Given that ECI compares predicted probabilities with observed probabilities that are in fact estimated, we advise bootstrapping to obtain confidence intervals or to test for a statistically significant difference between models. We always have to keep in mind that two models with the same ECI may have different calibration curves and thus a different type of miscalibration.

Since there is no one-to-one relationship between predicted and observed probabilities for the multiclass calibration plots (see Figure 3) we use smoothed curves to visualize the general relationship. If only the smoothed curves are shown (as we did in Figures 4-5 for clarity), information is lost. Nevertheless, we believe these smoothed relationship is most important in the assessment of calibration performance. Yet if this relationship is good, meaning that it is close to the diagonal line, it is still possible that predicted and observed probabilities are away from the diagonal line for many patients. So far, it is our experience that it is very difficult to develop a model where, the result for every patient in a validation dataset is close to the diagonal line. Further research on this issue is needed.

5.2 Tuning: logistic regression versus machine learning

The tuning of risk prediction models is an important issue [26]. For logistic regression, variable selection, non-linear effects and interaction terms can be considered as tuning [26]. We have not investigated the inclusion of non-linear effects or interactions, although non-linear effects are probably present [58]. In real clinical applications such tuning is advised, for example by modeling continuous predictors using spline functions, however the amount of tuning should depend on the sample size that is available [58]. Machine learning approaches have different tuning parameters (e.g. number of neighbors for kNN and number of considered variables in each split for RF) [26]. The choice of optimization method for these hyperparameters may influence model performance [31]. As expected, we noticed in our study that considering logloss as an optimization method for tuning resulted in better calibration compared with accuracy. We acknowledge that other optimization methods can be considered and may yield different model performance [31].

5.3 Choice of algorithms

In the literature, many machine-learning models are available and the optimal hyperparameter values of these models can be obtained using different optimization techniques. Different methods can be used to create probabilities, certainly for multiclass outcomes [25, 28]. This results in a plethora of possible algorithms to develop binary and multiclass risk models. In this paper we considered some approaches as an illustration of the multiclass calibration tools that may be used. We did not attempt to compare algorithms with respect to their capacity to produce calibrated risk estimates. The latter would be of interest, but to do this we believe that a large benchmark study would be required to compare a large battery of algorithms on multiple datasets.

5.4 Further practical limitations

We compared different degrees of freedom for the vector splines in the vector spline MLR analysis, and concluded that 2 degrees of freedom were sufficient [15]. Procedures for automatic selection of the level of smoothing for the calibration plots, for example by using generalized cross validation, would be desirable [59].

We randomly divided the data into development and validation data, with stratification for the multiclass outcome. This random split allowed the construction of calibration plots, but the results may depend on the split. However we expect that the dependency is limited due to the large sample size. Also, in real applications it is more interesting to use an external validation dataset, for example by using new data collected later in time or data from different hospitals. An advantage of using such a random split is that it should lead to few problems with mean calibration, because the development and validation are two random samples from an identical population. Miscalibration in our case study was more likely to be related to over-fitting.

Another limitation is that we decided to use single imputation for CA125. This approach ignores the fact that we are uncertain about whether the imputed values are correct [42]. An alternative to deal with the issue of uncertainty would be multiple imputation [42]. However, since the aim of this study was not the development of risk prediction models for clinical practice, we made a pragmatic decision to use single imputation.

5.5 General conclusions

The generic recalibration framework is an interesting approach to visualizing the reliability of predicted risks and quantify lack of calibration. The estimated calibration index (ECI) is easy for comparing models, but is less informative and interpretable than calibration plots.

Acknowledgments

Kirsten Van Hoorde is supported by a PhD grant of the Flanders' Agency for Innovation by Science and Technology (IWT Vlaanderen). Dirk Timmerman is senior clinical investigator of the Research Foundation – Flanders (FWO). Research supported by the Flemish Government (FWO project G049312N, IWT project TBM 070706-IOTA3, iMinds Medical Information Technologies SBO 2014), Research Council KU Leuven (GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC)); Belgian Federal Science Policy Office (IUAP P7/19/ DYSCO, `Dynamical systems, control and optimization', 2012-2017); and European Research Council (ERC Advanced Grant: BIOTENSORS (n° 339804)). Tom Bourne is supported by the NIHR Biomedical Research Center based at Imperial College Healthcare NHS Trust and Imperial College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, United States of America, 2009.

[2] König I, Malley J, Weimar C, Diener HC, Ziegler A. Practical experiences on the necessity of external validation. *Statistics in Medicine* 2007; **26**:5499–5511. DOI: 10.1002/sim.3069.

[3] Altman D, Vergouwe Y, Royston P, Moons K. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; **338**:b605. DOI: 10.1136/bmj.b605.

[4] Peek N, Abu-Hanna A. Clinical prognostic methods: trends and developments. *Journal of Biomedical Informatics* 2014; **48**:1–4. DOI: 10.1016/j.jbi.2014.02.016.

[5] Toll D, Janssen K, Vergouwe Y, Moons K. Validation, updating and impact of clinical prediction rules: A review. *Journal of Clinical Epidemiology* 2008; **61**:1085–1094. DOI: 10.1016/j.jclinepi.2008.04.008.

[6] Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. *Journal of biomedical informatics* 2005; **38**(5):367–375.

[7] Van Calster B, Vickers AJ. Calibration of risk prediction models impact on decision-analytic performance. *Medical Decision Making* 2014; :0272989X14547233.

[8] Janes H, Pepe MS, Huang Y. A framework for evaluating markers used to select patient treatment. *Medical Decision Making* 2014; **34**(2):159–167.

[9] Helfand M. Shared decision making, decision aids, and risk communication. *Medical Decision Making* 2007; .

[10] Brinkman S, Abu-Hanna A, van der Veen A, de Jonge E, de Keizer NF. A comparison of the performance of a model based on administrative data and a model based on clinical data: Effect of severity of illness on standardized mortality ratios of intensive care units*. *Critical Care Medicine* 2012; **40**(2):373–378. DOI: 10.1097/CCM.0b013e318232d7b0.

[11] Simon R. The use of genomics in clinical trial design. *Clinical Cancer Research* 2008; **14**(19):5984–5993.

[12] Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine* 1975; **293**(5):229–234.

[13] Taktak AF, Eleuteri A, Lake SP, Fisher AC. A web-based tool for the assessment of discrimination and calibration properties of prognostic models.

Computers in Biology and Medicine 2008; **38**:785–791. DOI: 10.1016/j.compbiomed.2008.04.005.

[14] Austin PC, Steyerberg EW. Graphical assessment of interal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine* 2014; **33**(3):517–535. DOI: 101002/sim.5941.

[15] Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Assessing calibration of multinomial risk prediction models. *Statistics in Medicine* 2014; **33**(15):2585–2596. DOI: 10.1002/sim.6114. http://dx.doi.org/-10.1002/sim.6114.

[16] Condous G, Van Calster B, Kirk E, Haider Z, Timmerman D, Van Huffel S, Bourne T. Prediction of ectopic pregnancy in women with a pregnancy of unknown location. *Ultrasound in Obstetrics & Gynecology* 2007; **29**(6):680–687. DOI: 10.1002/uog.4015.

[17] Djavan B, Remzi M, Zlotta A, Seitz C, Snow P, Marberger M. Novel artificial neural network for early detection of prostate cancer. *Journal of Clinical Oncology* 2002; **20**(4):921–929.

[18] Kattan MW. Comparison of cox regression with other methods for determining prediction models and nomograms. *The Journal of Urology* 2003; **170**(6):S6–S10. DOI: 10.1097/01.ju.0000094764.56269.2d.

[19] Van Esbroeck A, Rubinfeld I, Hall B, Syed Z. Quantifying surgical complexity with machine learning: Looking beyond patient factors to improve surgical models. *Surgery* 2014; DOI: 10.1016/j.surg.2014.04.034.

[20] Klement RJ, Allgäuer M, Appold S, Dieckmann K, Ernst I, Ganswindt U, Holy R, Nestle U, Nevinny-Stickel M, Semrau S, Sterzing F, Wittig A, Andratschke N, Guckenberger M. Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer. *International Journal of Radiation Oncology* Biology* Physics* 2014; **88**(3):732–738. DOI: 10.1016/j.ijrobp.2013.11.216.

[21] Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks* 2006; **19**(4):408–415. DOI: 10.1016/j.neunet.2005.10.007.

[22] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics* 2006; **2**:59.

[23] Sajda P. Machine learning for detection and diagnosis of disease. *Annual Review of Biomedical Engineering* 2006; **8**:537–565. DOI: 10.1146/annurev.bioeng.8.061505.095802.

[24] Kruppa J, Ziegler A, König IR. Risk estimation and risk prediction using machine-learning methods. *Human Genetics* 2012; **131**(10):1639–1654. DOI: 10.1007/s00439-012-1194-y.

[25] Van Calster B, Luts J, Suykens JA, Condous G, Bourne T, Timmerman D, Van Huffel S. Comparing methods for multi-class probabilities in medical decision making using ls-svms and kernel logistic regression. In *Artificial Neural Networks - ICANN 2007 lecture notes in computer science*, Marques de Sá J, Alexandre L, Duch W, Mandic D, eds. 139–148.

[26] Kruppa J, Liu Y, Biau G, Kohler M, König IR, Malley JD, Ziegler A. Probability estimation with machine learning methods for dichotomous and multi-category outcome: Theory. *Biometrical Journal* 2014; **56**(4):534–563. DOI: 10.1002/bimj.201300068.

[27] Kruppa J, Liu Y, Diener HC, Holste T, Weimar C, König IR, Ziegler A. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications. *Biometrical Journal* 2014; **56**(4):564–583. DOI: 10.1002/bimj.201300077.

[28] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eight International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, 694–699.

[29] Malley J, Kruppa J, Dasgupta A, Malley K, Ziegler A. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods of Information in Medicine* 2012; **51**(1):74. DOI: 10.3414/ME00-01-0052.

[30] Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association* 2012; **19**(2):263–274. DOI: 10.1136/amiajnl-2011-000291.

[31] Matheny ME, Resnic FS, Arora N, Ohno-Machado L. Effects of svm parameter optimization on discrimination and calibration for post-procedural pci mortality. *Journal of Biomedical Informatics* 2007; **40**(6):688–697. DOI: 10.1016/j.jbi.2007.05.008.

[32] Jiang X, Menon A, Wang S, Kim J, Ohno-Machado L. Doubly optimized calibrated support vector machine (doc-svm): An algorithm for joint optimization of

discrimination and calibration. *PloS One* 2012; **7**(11):e48823. DOI: 10.1371/journal.pone.0048823.

[33] Agresti A. *Categorical Data Analysis*. Wiley series, United States of America, 2002.

[34] Yee T, Wild C. Vector generalized additive models. *Journal of the Royal Statistical Society Series B (Methodological)* 1996; **58**(3):481–493.

[35] Yee TW. Vector splines and other vector smoothers. In *COMPSTAT: Proceedings in Computational Statistics*, Betlehem JG, van der Heijden PGM, eds. Springer, 529–534.

[36] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology* 2010; **21**(1):128–138. DOI: 10.1097/EDE.0b013e3181c30fb2.

[37] Timmerman D, Testa A, Bourne T, Ferrazzi E, Ameye L, Konstantinovic M, Van Calster B, Collins W, Vergote I, Van Huffel S, Valentin L. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group. *Journal of Clinical Oncology* 2005; **23**:8794–8801. DOI: 10.1200/JCO.2005.01.7632.

[38] Timmerman D, Van Calster B, Testa A, Guerriero S, Fischerova D, Lissoni A, Van Holsbeke C, Fruscio R, Czekierdowski A, Jurkovic D, Savelli L, Vergote I, Bourne T, Van Huffel S, Valentin L. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the iota group. *Ultrasound in Obstetrics & Gynecology* 2010; **36**:226–234. DOI: 10.1002/uog.7636.

[39] Van Holsbeke C, Van Calster B, Testa A, Domali E, Lu C, Van Huffel S, Valentin L, Timmerman D. Prospective interval validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. *Clinical Cancer Research* 2009; **15**:648–691. DOI: 10.1158/1078-0432.CCR-08-0113.

[40] Van Calster B, Van Hoorde K, Valentin L, Testa AC, Fischerova D, Van Holsbeke C, Savelli L, Franchi D, Epstein E, Kaijser J, Van Belle V, Czekierdowski A, Guerriero S, Fruscio R, Lanzani C, Scala F, Bourne T, Timmerman D, International Ovarian Tumour Analysis (IOTA) group. Evaluating the risk of ovarian cancer prior to surgery using the adnex risk model: diagnostic study to differentiate

between benign, borderline, stage i invasive, advanced stage invasive, and secondary metastatic tumours. *BMJ* 2014; **349**:g5920. DOI: 10.1136/bmj.g5920.

[41] Schenker N, Taylor J. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis* 1996; **22**:425–446. DOI: 10.1016/0167-9473(95)00057-7.

[42] Sterne J, White I, Carlin J, Spratt M, Royston P, Kenward M, Wood A, Carpenter J. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**:b2393. DOI: 10.1136/bmj.b2393.

[43] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics* 2005; **38**(5):404–415.

[44] Van Calster B, Vergouwe Y, Van Belle V, Looman C, Timmerman D, Steyerberg E. Assessing the discriminative ability of risk models for more than two outcome categories: a perspective. *European Journal of Epidemiology* 2012; **27**(10):761–770. DOI: 10.1007/s10654-012-9733-3.

[45] Hand DJ, Till RJ. A simple generalization of the area under the roc curve for multiple class classification problems. *Machine Learning* 2001; **45**:171–186. DOI: 10.1023/A:1010920819831.

[46] Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the c statistic to nominal polytomous outcomes: The polytomous discrimination index. *Statistics in Medicine* 2012; **31**:2610–2626. DOI: 10.1002/sim.5321.

[47] Schölkopf B, Smola AJ. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[48] Steinbach M, Tan PN. knn: k-nearest neighbors. In *The Top Ten Algorithms in Data Mining*, Wu X, Kumar V, eds. Chapman & Hall/CRC, 2009; 151–162.

[49] Breiman L. Random forests. *Machine learning* 2001; **45**(1):5–32. DOI: 10.1023/A:1010933404324.

[50] Hand DJ. Naïve bayes. In *The Top Ten Algorithms in Data Mining*, Wu X, Kumar V, eds. Chapman & Hall/CRC, 2009; 163–178.

[51] Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science* 2003; **18**(1):104–117.

[52] Frank E, Kramer S. Ensembles of nested dichotomies for multi-class problems. In *Proceedings of the 21st International Conference on Machine Learning*. Banff, Canada, 39.

[53] Hastie T, Tibshirani R. Classification by pairwise coupling. *Annals of Statistics* 1998; **26**(2):451–471. DOI: 10.1214/aos/1028144844.

[54] Wu T, Lin C, Weng R. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 2004; **5**:975–1005.

[55] Rish I. An empirical study of the naive bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3. 41–46.

[56] Harrell F. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, UK, 2001.

[57] Hosmer D, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression models. *Statistics in Medicine* 1997; **16**:965–980. DOI: 10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O.

[58] Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. *Biometrical Journal* 2014; **56**(4):601–606. DOI: 10.1002/bimj.201300297.

[59] Golub GHG, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 1979; **21**(2):215–223. DOI: 10.1080/00401706.1979.10489751.

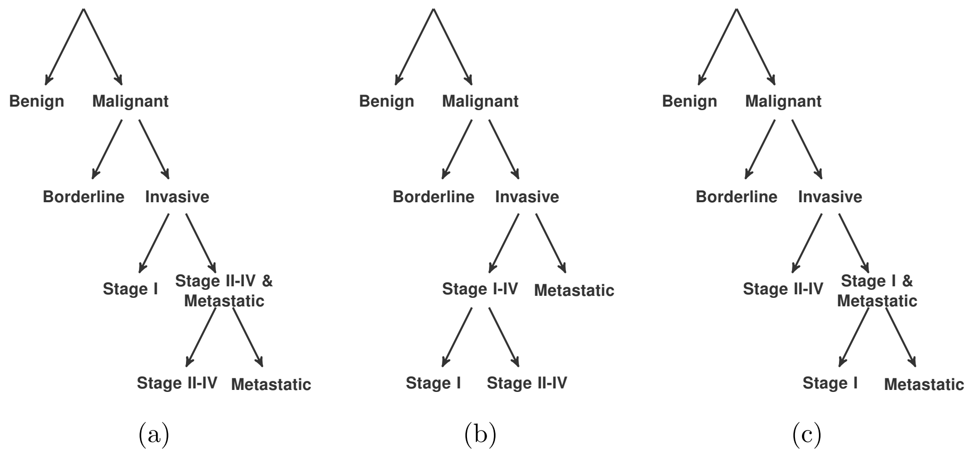
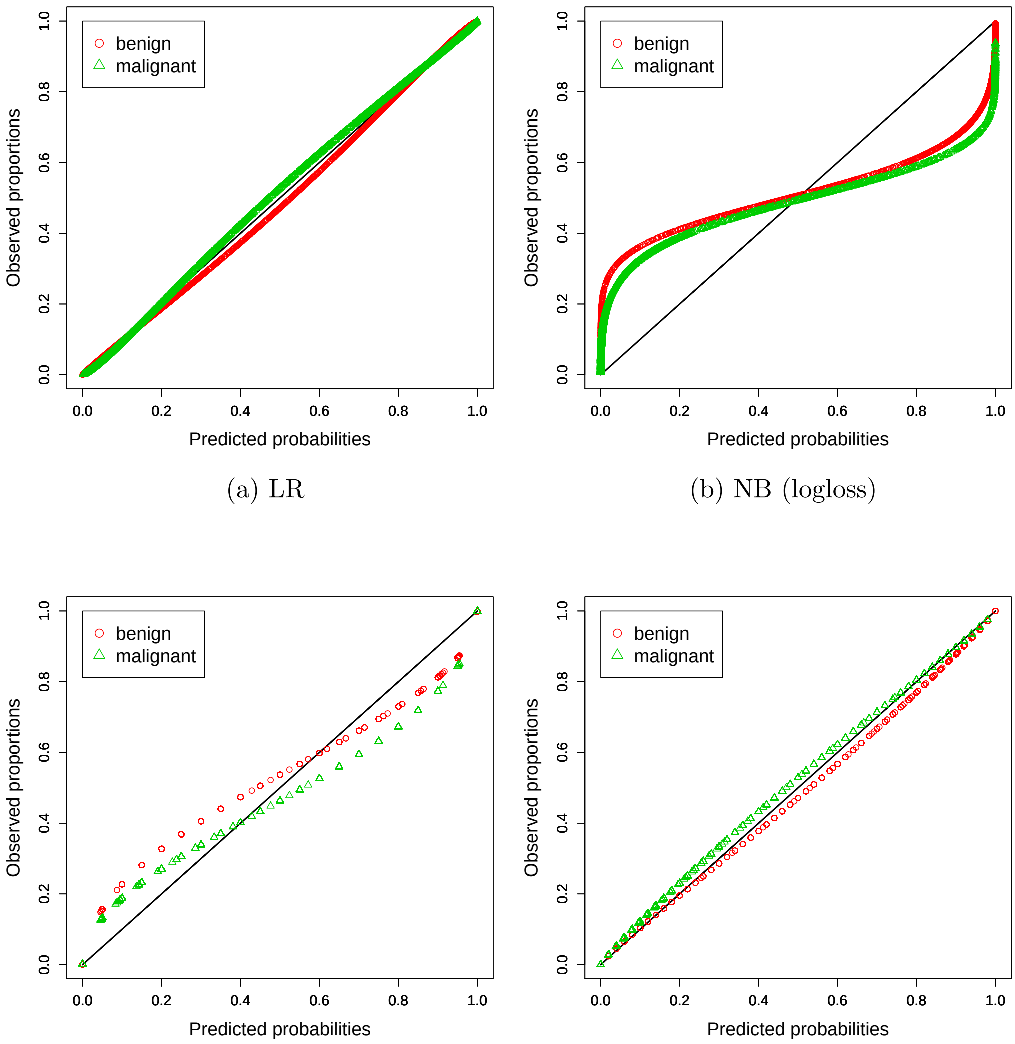


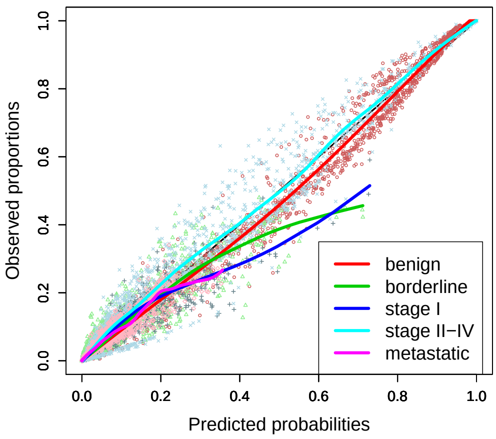
Figure 1: The three different considered sequential dichotomous models or binary trees.



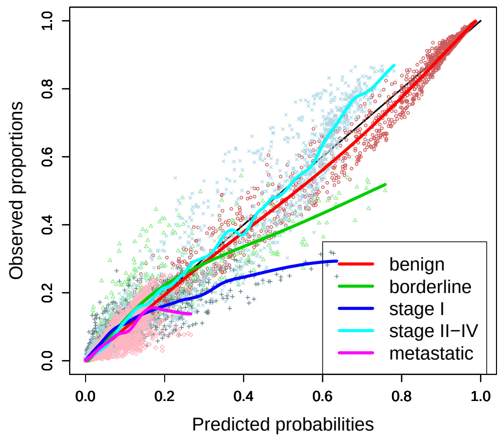
(a) LR (b) NB (logloss)

(c) KNN (accuracy) (d) KNN (logloss)

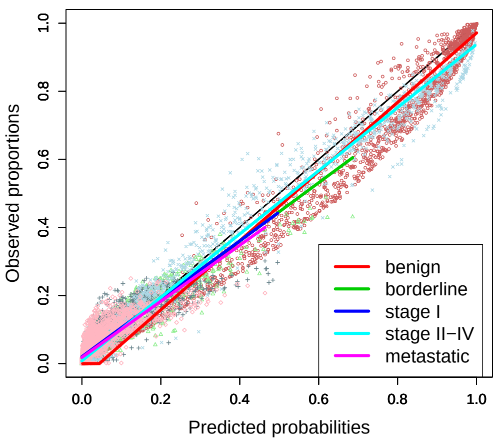
Figure 2: Calibration plot of the validation data for (a) dichotomous logistic regression, (b) binary naive Bayes, (c) k nearest neighbors with maximizing accuracy and (d) minimizing logloss as selection criterion.



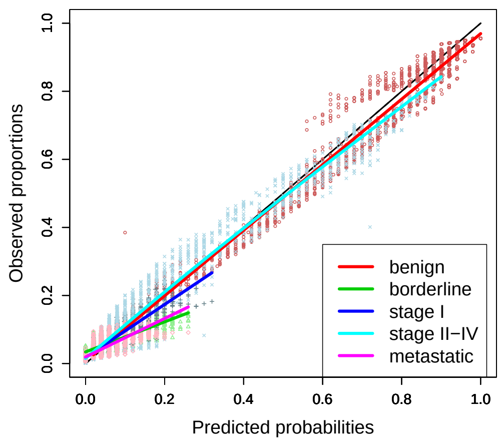
(a) LR-PC



(b) SVM-BT (logloss)

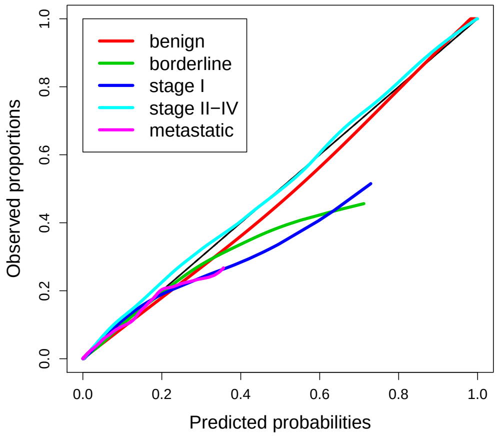


(c) RF

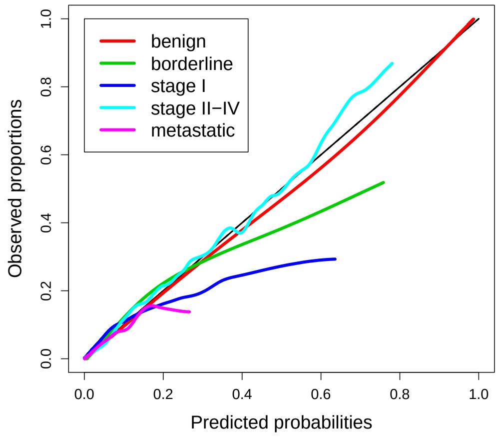


(d) kNN (logloss)

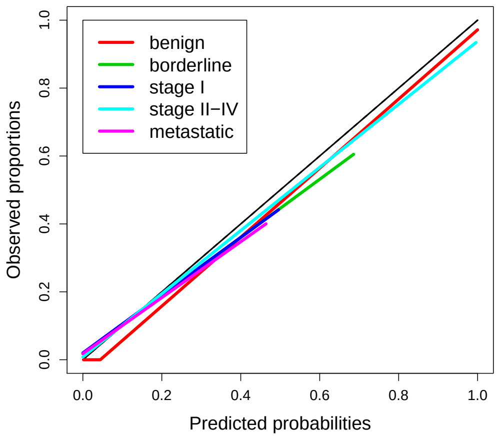
Figure 3: Multiclass calibration plot with smoother of the validation data for (a) pairwise coupling logistic regression model, (b) binary tree support vector machines with logloss as optimization method, (c) random forest and (d) k nearest neighbors with logloss as optimization method.



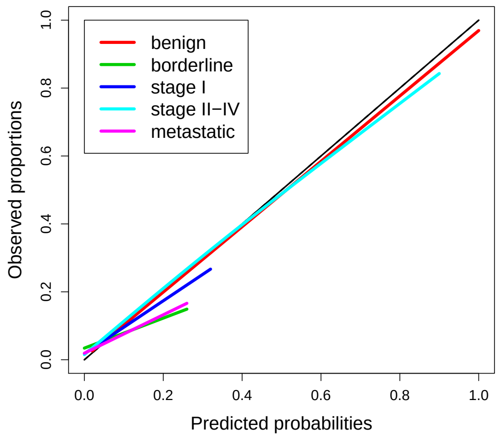
(a) LR-PC



(b) SVM-BT (logloss)

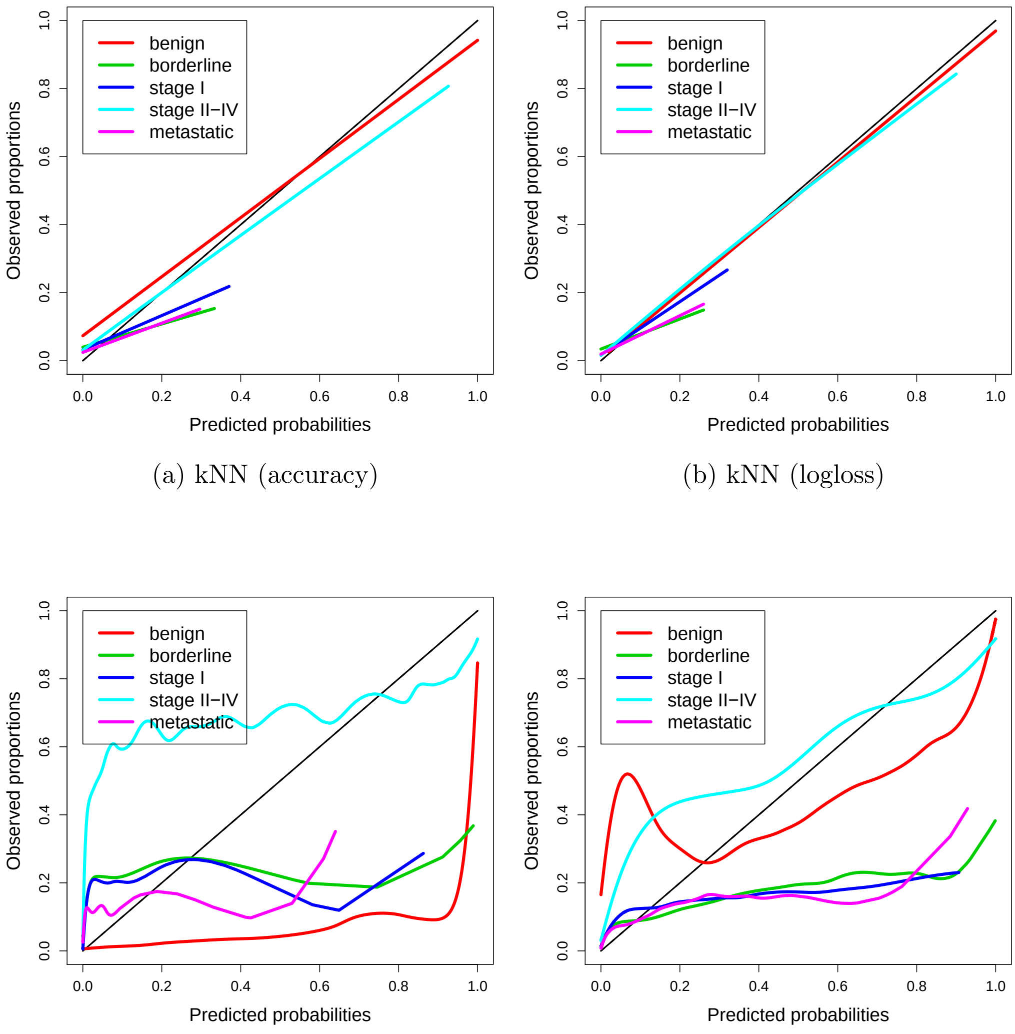


(c) RF



(d) kNN (logloss)

Figure 4: Smoothed multiclass calibration plot of the validation data for (a) pairwise coupling logistic regression model, (b) binary tree support vector machines with logloss as optimization method, (c) random forest and (d) k nearest neighbors with logloss as optimization method.



(a) kNN (accuracy)

(b) kNN (logloss)

(c) NB (accuracy)

(d) NB (logloss)

Figure 5: Smoothed multiclass calibration plot of the validation data for k nearest neighbors (a-b) and naive Bayes (c-d) with accuracy and logloss as optimization method.

Supplementary Table
Click here to download Table: calibration of multiclass risk prediction models - supplementary tables.docx

Table S1. Conditional pairwise AUCs for the multiclass prediction models for the validation data.

Model (selection criterion)	Be vs Bo	Be vs St*1*	Be vs St24	Be vs Met	Bo vs St1	Bo vs St24	Bo vs Met	St1 vs St24	St1 vs Met	St24 vs Met
MLR	0.887	0.917	0.964	0.950	0.702	0.881	0.893	0.773	0.722	0.684
LR-BT	0.887	0.920	0.967	0.949	0.724	0.881	0.883	0.782	0.724	0.704
LR-PC	0.883	0.919	0.966	0.950	0.710	0.881	0.886	0.780	0.730	0.698
SVM-BT (accuracy)	0.821	0.919	0.975	0.949	0.708	0.847	0.831	0.756	0.717	0.604
SVM-BT (logloss)	0.889	0.919	0.973	0.952	0.708	0.864	0.849	0.751	0.676	0.595
SVM-PC (accuracy)	0.710	0.887	0.973	0.937	0.733	0.895	0.898	0.746	0.744	0.594
SVM-PC (logloss)	0.701	0.896	0.974	0.928	0.730	0.895	0.888	0.746	0.724	0.608
kNN (accuracy)	0.683	0.784	0.945	0.797	0.589	0.799	0.622	0.750	0.517	0.669
kNN (logloss)	0.676	0.805	0.945	0.811	0.583	0.824	0.632	0.797	0.577	0.698
RF (accuracy)	0.880	0.919	0.983	0.929	0.748	0.912	0.864	0.793	0.676	0.695
RF (logloss)	0.880	0.919	0.983	0.929	0.748	0.912	0.864	0.793	0.676	0.695
NB (accuracy)	0.867	0.887	0.973	0.932	0.720	0.895	0.892	0.785	0.742	0.693
NB (logloss)	0.868	0.874	0.955	0.920	0.680	0.868	0.864	0.762	0.696	0.604
NSC (accuracy)	0.846	0.855	0.935	0.897	0.671	0.848	0.876	0.744	0.726	0.674
NSC (logloss)	0.844	0.859	0.937	0.898	0.679	0.850	0.880	0.743	0.728	0.674

Abbreviations: Be, benign; Bo, borderline; St1, Stage I; St24, Stage II-IV; Met, Metastatic.

Table S2. Mean calibration for the multiclass prediction models for the validation data.

Model (selection criterion)	Benign	Borderline	Stage I	Stage II-IV	Metastatic
MLR	-0.13	-0.10	-0.16	0.37	0.03
LR-BT	-0.25	-0.31	-0.08	0.77	-0.13
LR-PC	-0.25	-0.40	-0.08	0.73	0.00
SVM-BT (accuracy)	-0.94	-0.80	0.14	1.51	-0.19
SVM-BT (logloss)	-0.40	0.01	-0.23	0.70	-0.40
SVM-PC (accuracy)	2.13	-1.63	-1.10	0.89	-0.29
SVM-PC (logloss)	2.10	-1.67	-1.01	0.86	-0.27
kNN (accuracy)	-1.74	0.58	0.45	0.54	0.17
kNN (logloss)	-1.92	0.52	0.57	0.61	0.21
RF (accuracy)	-3.35	1.10	1.31	-0.38	1.31
RF (logloss)	-3.35	1.10	1.31	-0.38	1.31
NB (accuracy)	-22.60	4.90	5.49	8.44	3.77
NB (logloss)	-0.48	-4.00	-0.45	5.81	-0.88
NSC (accuracy)	-2.80	0.10	-0.36	2.89	0.16
NSC (logloss)	-3.52	0.20	-0.27	3.28	0.31

Appendices

Click here to download Supplementary Material: calibration of multiclass risk prediction models - appendix - 181214.docx